# The effect of Immigration on Labor Market Outcomes

Ruxin Chen

1 Data Collection

The IPUMS preserves U.S. census microdata from 1790 to 2010 and American Community Surveys (ACS) from 2000 to the present. The database contains large samples of individual demographic characteristics, geographical location, labour market outcomes that are sufficient for our study. We initially requested a data set consists of more than 3,800,000 observations from 2001 - 2016, which occupies about 4 GB of space. Variables include state, county, metropolitan status, age, sex, marital status, race, ethnicity, year of immigration, language spoken, education attainment, labor force status, employment status, occupation, class of worker, usual hours worked per week, place of work in terms of state and wage and salary income. We create some dummy variables to denote some of the above categorical data. Since we are interested in the changes of labor market outcomes, we drop the observations that are not in the labor force. Finally, we cluster the data by county and year, so that the categorical data are denoted as the proportion of population in each group. For example, a dummy variable "Female" is created in the raw data indicating the sex of the individual, in the clustered data, we calculate the mean of this variable, which is simply the proportion of female in the associated county and year. The clustered data have 4450 observations in total.

Next, we report the summary statistics for some selected variables of interest. The average proportion of immigrants is 12%. The distribution is right-skewed, which suggests that we have more counties with relative small percentage of immigrants, while very few counties have immigrants more than half of its entire population. During 2006-2015, the unemployment rate is about 7% with a standard error of 0.03%, the shape of the histogram is close to a normal distribution except that it appears to be a bit right-skewed. In our sample, the majority of the population (54%) entered in colleges, on average, 20% of the population obtained a college degree and 12% have advanced degrees.
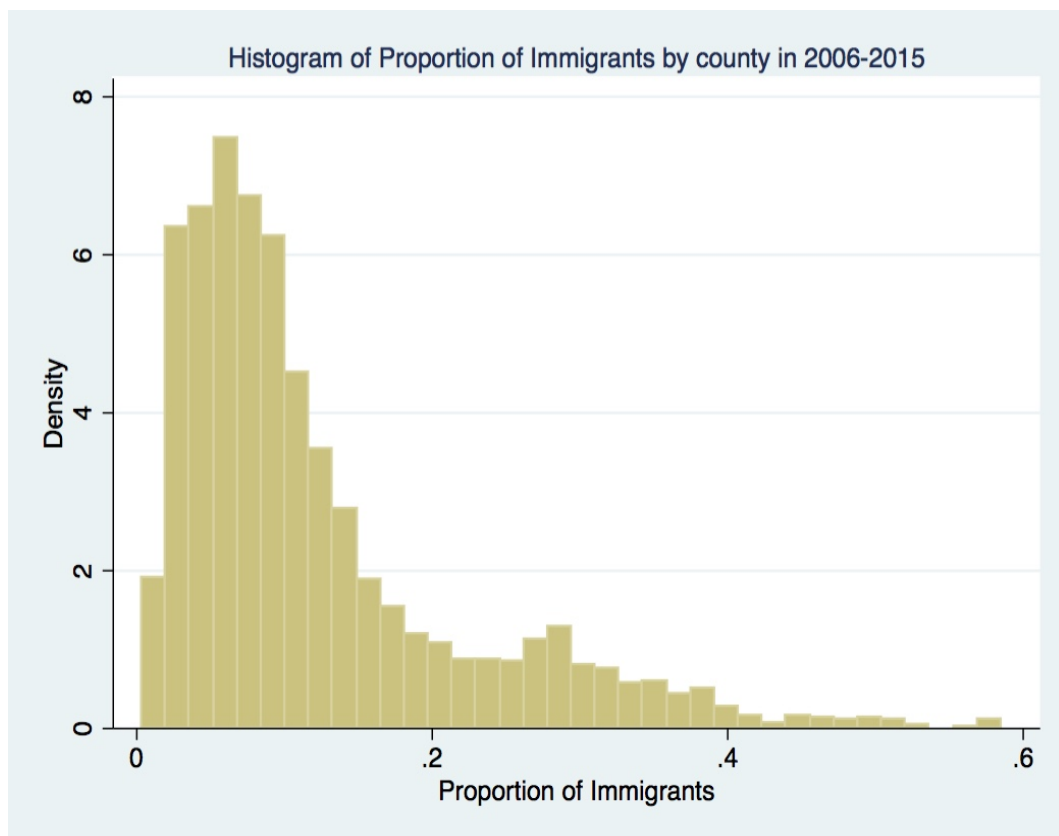
Histogram of Proportion of Immigrants by county in 2006-2015
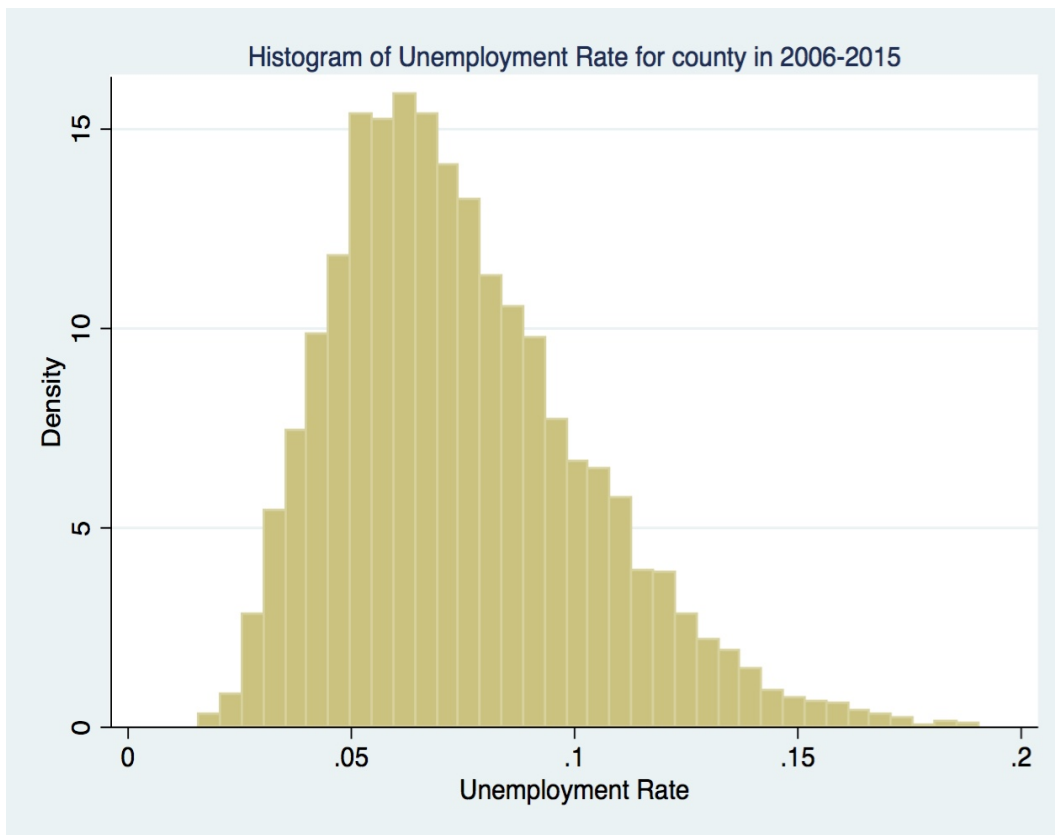
Table 1: Summary Statistics for Selected Variables

	mean	sd	min	max
Proportion of Immigrants	0.12	0.10	0.00	0.59
Unemployment Rate	0.07	0.03	0.02	0.19
Usual hours worked per week	37.56	1.57	33.04	45.85
Wage and salary income	40129.10	9712.91	21540.13	93110.58
Average Age	42.51	1.68	33.99	50.18
Proportion of Female	0.48	0.02	0.33	0.56
Proportion of White	0.76	0.29	0.00	1.00
Proportion of Black	0.09	0.11	0.00	0.69
Proportion of Hispanic	0.10	0.13	0.00	0.97
Proportion of Elementary	0.00	0.00	0.00	0.05
Proportion of Some College	0.54	0.14	0.11	0.81
Proportion of College Degree	0.20	0.06	0.07	0.42
Proportion of Advanced Degree	0.12	0.05	0.02	0.42
Proportion of No Education	0.01	0.01	0.00	0.04
Proportion of High School Degree	0.14	0.13	0.02	0.64
Observations	4450			

Table 2: Preliminary Result: OLS Regression For Unemployment Rate and Average Wage

	(1)	(2)
	Average Wage and Salary Income	Unemployment Rate
Proportion of Immigrants	15132.3***	0.0897***
	(1158.14)	(0.01)
Average Age	1360.0***	0.000234
	(41.14)	(0.00)
Proportion of Female	-105279.1***	0.255***
	(3724.24)	(0.02)
Proportion of White	-2485.1***	0.0173***
	(229.99)	(0.00)
Proportion of Black	2224.9**	0.0607***
	(697.92)	(0.00)
Proportion of Hispanic	-4701.5***	-0.00920*
	(913.66)	(0.00)
Proportion of Elementary	-29458.3	0.221
	(33758.85)	(0.17)
Proportion of Some College	13157.5	-0.219*
	(19126.67)	(0.10)
Proportion of College Degree	84810.3***	-0.433***
	(19009.72)	(0.10)
Proportion of Advanced Degree	92399.0***	-0.269**
	(18997.49)	(0.10)
Proportion of High School	11524.3	-0.290**
	(18945.63)	(0.10)
Constant	-2830.9	0.188
	(19118.59)	(0.10)
R-squared	0.810	0.417
N. of cases	4450	4450

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Preliminary Result: Fixed Effect Model

	(1) Unemployment Rate	(2) Unemployment Rate
Proportion of Immigrants	0.00521	0.0739***
	(0.01)	(0.01)
Average Age	-0.000667***	0.000123
	(0.00)	(0.00)
Proportion of Female	0.205***	0.228***
	(0.01)	(0.02)
Proportion of White	-0.0102	-0.00113
	(0.01)	(0.00)
Proportion of Black	0.0536***	0.0407***
	(0.01)	(0.01)
Proportion of Hispanic	0.0263***	-0.00783*
	(0.00)	(0.00)
Proportion of Elementary	-0.261*	0.138
	(0.12)	(0.14)
Proportion of Some College	-0.259***	-0.193*
	(0.07)	(0.08)
Proportion of College Degree	-0.380***	-0.376***
	(0.07)	(0.08)
Proportion of Advanced Degree	-0.296***	-0.227**
	(0.07)	(0.08)
Proportion of High School	-0.235***	-0.195*
	(0.07)	(0.08)
Year Fixed Effect	Yes	Yes
State Fixed Effect	Yes	No
Constant	0.265***	0.164
	(0.07)	(0.08)
R-squared	0.734	0.595
N. of cases	4450	4450

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2 Model

We specifies a fixed-effect model to describe the relationship between immigrants on US local labor market outcomes as follows:

$$y_{it} = \beta_0 + \beta_1 Immigrant_{it} + \gamma X_{it} + u_{it} \tag{1}$$

where y_{it} is the outcome of interests including the unemployment rate and the average wage for county i in year t. $Immigrant_{it}$ denotes the proportion of immigrants. X_{it} is a set of observable characteristics which contains the average age for the county, proportion of female, proportion of white/black population, proportion of Hispanics, and the education attainment level. Table 2 presents the OLS regression results. The coefficient for proportion of immigrants are significant for both average wage and unemployment rate. However, the preliminary result shows that after taking account into the state and time fixed effect, the effect of immigrants becomes insignificant to the labor market outcome. As mentioned in the literature review, a "zero" effect might be because the adverse effects of different subgroups just cancel out, and the major innovation of this paper is to investigate whether the effect is heterogeneous over the population, if so, identify these subgroups using a method called "honest" approach based on regression tree. This method uses the same model as specified in equation (1), the only difference is the vector X_{it} consists of dummy variables that divide our data into several subgroups, and they are not known prior to the estimation process. The estimation is not completed yet, hence results are not reported here.